Journal of Development Effectiveness

ISSN: 1943-9342 (Print) 1943-9407 (Online) Journal homepage: http://www.tandfonline.com/loi/rjde20

Assessing 'what works' in international development: meta-analysis for sophisticated dummies

Maren Duvendack , Jorge Garcia Hombrados , Richard Palmer-Jones & Hugh Waddington

To cite this article: Maren Duvendack , Jorge Garcia Hombrados , Richard Palmer-Jones & Hugh Waddington (2012) Assessing 'what works' in international development: meta-analysis for sophisticated dummies, Journal of Development Effectiveness, 4:3, 456-471, DOI: 10.1080/19439342.2012.710642

To link to this article: http://dx.doi.org/10.1080/19439342.2012.710642

Published online: 18 Sep 2012.

Journal of Development Effectiveness
Vol. 4, No. 3, September 2012, 456–471

Routledge
Taylor & Francis Group

Assessing 'what works' in international development: meta-analysis for sophisticated dummies

Maren Duvendack[a]*, Jorge Garcia Hombrados[b], Richard Palmer-Jones[c] and Hugh Waddington[b]

[a]*Overseas Development Institute, 111 Westminster Bridge Road, London SE1 7JD, UK;* [b]*International Initiative for Impact Evaluation (3ie), London, UK;* [c]*School of International Development, University of East Anglia, Norwich, UK*

Many studies of development interventions are individually unable to provide convincing conclusions because of low statistical significance, small size, limited geographical purview and so forth. Systematic reviews and meta-analysis are forms of research synthesis that combine studies of adequate methodological quality to produce more convincing conclusions. In the social sciences, study designs, types of analysis and methodological quality vary tremendously. Combining these studies for meta-analysis entails more demanding risk of bias assessments to ensure that only studies with largely appropriate methodological characteristics are included, and sensitivity analysis should be performed. In this article, we discuss assessing risk of bias and meta-analysis using such diverse studies.

Keywords: impact evaluation; systematic review; meta-analysis; effect size calculation

1. Introduction

The evidence-based policy movement has largely focused on generating knowledge of 'what works' from primary quantitative studies attributing outcomes to particular interventions. Research designs, which are now commonly used for quantitative attribution of impacts of international development interventions, include experimental designs or randomised control trials (RCTs) (for example, Duflo *et al*. 2007) and quasi-experimental designs using statistical techniques (for example, Ravallion 2007). Synthesis of research on 'what works' in international development using systematic review and meta-analysis is now emerging.[1]

Research synthesis comprises two components: measures to control bias and quantitative statistical synthesis, termed meta-analysis (Chalmers *et al*. 2002, p. 16). Meta-analysis was formally developed by Smith and Glass (1977), and has subsequently become an important and popular method of research synthesis. Meta-analysis is 'the statistical combination of results from two or more separate studies' (Green *et al*. 2011); it has been widely used, particularly in the medical sciences, to synthesise the results of empirical studies of an intervention that addresses a common problem, using effect size as a measure of impact.

Meta-analysis is only possible for studies that can be meaningfully compared on a conceptual level. This means that similar variable constructs and relationships are used

*Corresponding author. Email: m.duvendack@odi.org.uk

ISSN 1943-9342 print/ISSN 1943-9407 online

http://dx.doi.org/10.1080/19439342.2012.710642
http://www.tandfonline.com

and they need to follow similar statistical approaches (Lipsey and Wilson 2001). In the traditional use of meta-analysis, most of the studies employ experimental designs whose quality and risk of bias can be assessed relatively easily (Coalition for Evidence-Based Policy 2010) and results are combined in fairly straightforward ways (Borenstein *et al*. 2009). Synthesis of quasi-experimental and observational studies is seen as more problematic; the studies are diverse, giving rise, for example to the so-called 'apples and oranges' problem, among others, for meta-analyses (Eysenck 1984, Sharpe 1997, Lipsey and Wilson 2001), and methodological quality is often hard to assess. This situation implies that risk of bias assessment prior to meta-analysis will be more demanding than where there are many studies based on experimental designs – although this is also more difficult for social and economic evaluations – or large epidemiological studies. Studies that are methodologically flawed or of low quality should not be included in meta-analysis since this could adversely affect the overall results (Slavin 1986), especially if they are likely to suffer common biases related to researcher allegiance, or institutional affiliation with funding, implementing or advocacy institutions. However, there is a debate among researchers as to what constitutes high methodological quality, whether statistical methods can control for biases, and how best the risk of bias should be assessed; these debates together with their implications for meta-analysis motivate our article.

We present the particular challenges of quality assessment and meta-analysis of diverse studies with variable methodological quality. We provide an overview of the factors affecting risk of bias in causal attribution for quasi-experimental designs and discuss an approach to assessing these risks. We then discuss the issues of effect size computation and the main challenges and limitations of meta-analysis of such a broad range of studies.

2. Risk of bias assessment in quasi-experimental designs

The 'quasi-experimental' designation covers a heterogeneous range of approaches. When programme assignment rules are observed and external to participants (as they are with RCTs), it is possible to model participation credibly using fairly simple statistical methods. Indeed, this is the principle behind regression discontinuity designs (RDDs) and 'natural' experiments (Shadish *et al*. 2002). Where assignment rules are not observed, as in the majority of quasi-experimental situations, other statistical methods are required, such as propensity score matching (PSM), difference-in-differences (DID) and instrumental variables (IV) regression (Ravallion 2007). The extent to which these other methods are able to control for both observable and unobservable sources of selection bias depends largely on the quality of the specifications and data used. Quasi-experimental designs are therefore more reliant on theoretical assumptions, which are harder than experimental approaches to assess 'objectively'. This is one reason why their inclusion in policy-relevant meta-analyses has been highly controversial.[2]

The empirical literature has shown that, for individual studies, weak methodological designs can lead to severe risks of bias in causal attribution, whereas well-conducted studies that carefully model participation (including IV, RDD and matching) can yield the same results as RCTs at an individual study level (Cook *et al*. 2008, Hansen *et al*. 2011). When comparing average differences across multiple randomised and non-randomised studies using meta-analysis, the evidence is mixed. One prominent article found significant differences in estimated effects between 12 replicated randomised and non-randomised studies (Glazerman *et al*. 2003), whereas others have suggested that the differences in results are almost zero overall, whether or not the individual study differences are themselves zero (Lipsey and Wilson 1993, Heinsman and Shadish 1996). Hansen *et al*. (2011) note, based

on an assessment of within-study comparisons, that appropriate knowledge of the participation decision process is key to estimating unbiased effects. It is therefore important to evaluate each study carefully prior to synthesis.

Risk of bias tools address and test the specific assumptions underpinning the validity of causal attribution methods. Any meta-analysis, whether it includes only RCTs or broader studies, should attempt to account for risks of biased effects in sensitivity analysis. In a recent systematic review, Duvendack *et al*. (2011) proposed a framework that combines both research design and statistical analysis to screen studies (Table 1). The table presents RCTs at one end of the design spectrum[3] and cross-section designs at the other, indicating that the methods of analysis have a significant potential to control the potential biases of quasi-experimental designs.[4]

Assessing risk of bias usually requires comprehensive evaluation of all criteria, which undermine estimation of causal effects, including selection bias, confounding, spill-overs and reporting biases (for example, CEBP 2010). Although the validity of quasi-experimental designs rests on qualitatively different criteria than randomised studies, the criteria used to evaluate them should be equivalent. The main differences when assessing risk of bias in quasi-experimental designs lie in the assessment of counterfactual identification, as well as a more comprehensive assessment of analysis reporting.[5]

The criteria on which sources of bias are evaluated in experimental studies are well known, subject to relatively objective assessment, and can be implemented by fairly non-specialist researchers (for example, those promulgated in Green *et al*. 2011 in the context of medicine, and in CEBP 2010 for social experiments). These focus on assessment of the randomisation process and the factors that might invalidate group equivalence, including non-random attrition, confounding, researcher allegiance and the differential performance of groups being monitored.

Bias assessment in quasi-experiments is complicated by the nature of the validity assumptions, for example unconfoundedness or exogeneity (Morgan and Winship 2007). There are a large number of existing tools to assess risk of bias, many of which enable integrated assessment of experimental and quasi-experimental designs[6]. The tools mostly rely on the assessment of group comparability in terms of observable covariates. Although some of the tools include vague questions about statistical validity, none provide further guidance to assess selection (and placement) bias and statistical analysis comprehensively. Operationalisation of existing tools to assess quasi-experimental designs used in development (including RDD, IV, PSM, DID) may therefore lead to simplistic and inappropriate study classifications. Adequate assessment of selection bias in quasi-experiments requires

Table 1. Potential risk of bias in quasi-experimental designs.

	Statistical methods of analysis		
Research design	DID, PSM, IV, RDD	Multivariate (or bivariate with covariate means tests)	Tabulation
RCT	Low	Low	Low–Medium
Natural experiment	Low	Low	Low–Medium
Pipeline	Low–Medium	Medium–high	High
Panel	Low–Medium	N/A	High
Cross section	Low–Medium	High	High

Source: Adapted from Duvendack *et al*. (2011).

analysis of the methods of counterfactual identification to address selection bias (that is not just whether the study used random allocation), among other factors including file-drawer effects and the use of appropriate specification tests.

3. Effect size and meta-analysis: challenges and limitations

When many studies of a topic have been undertaken, it is generally supposed that confidence in the conclusions can be enhanced by considering all the relevant analyses and pooling their results and this is what research synthesis aims to achieve (Chalmers *et al*. 2002). Furthermore, there has been much criticism of the traditional focus on significance testing (that there was a probability of less than say 5% that the estimate obtained could have occurred in the population from which the sample was obtained if the true relationship was zero; Rosenthal 1991) as the main way in which to assess the effect of interventions. Since the early days of modern statistical analysis, often dated to Fisher's (1925) publication, there have been critics of 'null hypothesis significance testing' (NSHT; see Nickerson 2000).[7] Critics claim that statistical significance is not the same as substantive significance and that exclusive focus on significance of individual studies fails to provide cumulation of results since it is not considered legitimate to consider prior findings in evaluating the current one. Thus, a common exposition contrasts a nearly significant (at 95% confidence) but substantively large effect, perhaps from a study with few observations, which is rejected and plays no further role in decision-making, and with another that shows a highly significant but substantively small effect, probably from a study with a large number of observations. Rejecting the former, it is argued that it may result in significant welfare losses, although accepting the latter yields few benefits. Surely, there is something wrong with such a decision-making process.

This dilemma can be appreciated by considering the two traditional types of error, which statistical testing aims to avoid – types 1 and 2. A type 1 error is when one rejects a true null (no effect of the intervention) hypothesis and, it is implied, acts on the (erroneous) assumption that the intervention has a meaningful effect (errors of commission). Type 2 is when one rejects a true hypothesis that the intervention has a meaningful effect, thereby foregoing the benefits that would follow from implementing the intervention more widely. Traditional NHST gives priority to avoiding type 1 errors; the 'new statistics' gives priority to avoiding type 2 errors (errors of omission) (Cumming 2012).

Two concepts have been developed to address the dominance of NHST – effect size (statistical) and power (see Field 2009, section 2.6 for an introduction). Statistical power is the ability to identify an effect if it is there; the effect size is the estimated substantive difference the intervention makes. Effect size on average should not vary with sample size, but power does; indeed, power varies with the (true) effect size, the variability of the effect and the sample size. Power calculations are used to decide the sample size required to detect, with a given degree of confidence, an effect of a given size subject to the assumed variability of the effect[8]; thus, for a given underlying variability, a large effect size can be detected as statistically significant with a smaller sample size and so on. The problem, it is often claimed, is that the sample size in many studies is too small to produce statistically significant results, unless the true effect is very large. However, suppose many under-powered studies give effects in the same direction (positive or negative) of varying size. One can either choose to reject all (or most) of them as failing to show a significant effect or can combine the studies and see if together they suggest a significant effect. This is what meta-analysis attempts to do (see Lipsey and Wilson 2001, Borenstein *et al*. 2009, Ellis 2010, Green *et al*. 2011).

In practice, most social science studies of a common intervention in development provide impact estimates in diverse metrics and statistics. For meta-analysis, impacts estimated in different metrics are converted into effect sizes, which are then pooled together along with covariates characterising each study which may be used to examine heterogeneity in effects. Furthermore, social science studies in development involve somewhat different interventions, use different research methods, assess different indicators of outcomes and occur in different contexts (time, place and social group), all of which may be expected to influence the specific conclusions reached; there may well be considerable heterogeneity in the precise relationships between intervention, context and outcome variables. Given the controversial nature of research synthesis using effect sizes computed from such diverse studies, we briefly outline some of the methods and the issues involved.

In order to synthesise findings from multiple studies, it is necessary to homogenise the impact results of included studies, putting them on a common scale to ensure comparability[9]. This process of standardisation is mainly achieved through the computation of effect sizes, which are defined (Borenstein *et al*. 2009) as a value that reflects the magnitude of the treatment effect or the strength of the association between two variables and is comparable across studies using different scales and estimation methods. Effect sizes are therefore the basic observation unit in meta-analysis. A good effect size needs to be comparable across studies and only reflect the magnitude of effect for each study, and not other factors such as sample size or the influence of confounding factors.

There are three types of effect size scales: *d*-values and *r*-values for continuous variables, and odds or risk ratios for categorical variables. The *d*-values are differences in outcomes standardised by their variability measured by their standard deviation or a close approximation, and are consequently similar to *z*-scores. The *r*-values are correlation coefficients, which range from -1 to $+1$. Odds and risk ratios are relevant to categorical data such as the occurrence of a disease, achievement of an educational status or survival. Other common metrics of outcome are scales or ordinal values, which can be transformed into incidence density ratios with similar interpretation to risk ratios. There are many commonly described transformations among different effect sizes (Lipsey and Wilson 2001, Ellis 2010), but it is important to recognise that the outcome of transformations do not always agree (see McGrath and Meyer 2006).

3.1. Experimental data

Most common methods of calculation of effect size are based on experimental research designs, such as clinical trials where the outcomes are either continuous variables like marks on an exam or dichotomous categorical variables like survival, infection and so on. The treatment indicator is generally also a dichotomous variable (treated/not treated), or sometimes an ordinal or continuous variable (level of treatment). Randomisation of allocations of units to treatment and control groups allows estimation of treatment effects by simple comparisons of these groups using either continuous or categorical variables representing outcomes.[10] The established tools for effect size calculation based on this epistemological context readily compute equivalent effect sizes from different reported statistics.[11] These 'zero-order' methods of analysis are generally equivalent. Thus, a *t*-test, analysis of variance and regression with dummy variables for treatment are algebraically equivalent, resulting in the same estimates of effect size and statistical significance (Lipsey and Wilson 2001). It is also important to bear in mind that the effect sizes and their confidence intervals derived from zero-order analyses are reliable only when certain conditions are met, specifically normality and homoscedasticity and independence (see Grissom and

Kim 2012 and Wilcox 2012 for an introduction). Effect size estimates can be biased by non-normality and heteroscedasticity (Wilcox 2008), although these statistics are often not reported. Under these circumstances, there is a trade-off between the risks of bias in effect size estimate because the statistics exhibit heteroscedasticity, and the costs of contacting the original authors to get these statistics or producing them by exact replication of the original study.

3.2. Quasi-experimental data

With non-experimental information, or experimental information that has to take account of potential confounding or covariate variables, the calculation of effect size is considerably less well developed. Most, but not all, of the studies with which we are concerned in development studies use some sort of multiple regression analysis (ordinary least squares (OLS), IV, logit, probit, tobit and so on). As is well known, regression coefficients can change dramatically when additional control variables are introduced into the estimation. The problem is that few studies are likely to control for all confounders, and studies will generally not include estimates using the same set of potential confounders even if the same underlying model of impact is being employed. Hence, the resulting impact estimates may not be strictly comparable. Further, in studies where unobservable factors are thought potentially to confound impact estimates (ability, motivation, degrees of risk aversion and so on), all studies will suffer from common biases. It is these considerations that make pooling effect sizes estimated by multivariate methods from diverse contexts problematic and potentially liable to pervasive bias where low-quality studies are included in the meta-analysis.

The estimation of effect size to be pooled with those from other similar studies seems to be much less well developed and more controversial for quasi-experimental studies. This is partly because such studies are 'partial' or non-zero-order. In these cases, the specifications of the estimation can have significant implications for effect size estimations (Jones 1992, Keef and Roberts 2004), and in some cases, such estimations may indeed be inappropriate (Colliver *et al*. 2008).

3.3. Multiple treatments, multiple outcomes and multiple methods

As mentioned above, meta-analysis aims to combine results of studies to gain greater confidence in conclusions than would be warranted by the individual studies taken separately. The presumption is that this is generally seen as legitimate when there is high homogeneity with respect to treatment, context and outcome; that is the studies are testing a common hypothesis in a comparable way. A meta-analysis needs to establish that it is indeed the case.

Petticrew and Roberts (2006) emphasise that 'Meta-analysis should only be applied when a series of studies has been identified for review that address an identical conceptual hypothesis' (p. 205, box 6.13). The question then arises as to what 'identical' means and how identicality can be established. A particular case arises when, although addressing an issue with an 'identical' conceptual framing, some studies treat intermediate or proxy variables as outcomes rather than indicators of ultimate outcomes (by which we mean variables which represent human well-being or freedoms; Sen 1999). In development studies, we often have results that can be considered 'intermediate' or 'instrumental' in attaining the true outcome, whether or not we are able to measure accurately the latter using quantitative methods. Thus, some studies compare treatment A with 'intermediate' outcomes B, and they, or others compare B with 'final' outcomes C.[12] It may be possible to indirectly

link A with C through their common links with B. A further relevant characteristic of the mainstream literature and application of meta-analysis (in medicine for example) is therefore that both treatment and outcome are equivalent in the relevant way, and generally fairly standardised in the field; they should also employ common methods of analysis and reporting.

However, it is not uncommon that measures of outcomes vary and there are many suggestions as to how to combine studies that use different outcome indicators, or the results within a single study that report estimates of effect for different metrics, and, or different interventions, or sub-groups, provided they all correspond to the same construct of the 'true' outcome (Hedges and Olkin 1985, Cooper and Hedges 1994, Sutton *et al*. 1998). This means that one is dealing with strictly comparable treatments and outcomes; or with comparable treatments with multiple indicators of outcomes.

There are also cases where treatments may be considered conceptually identical but superficially differ. However, Raudenbusch (2009, p. 296) refers to 'definitions of treatment' as a characteristic of studies that might account for why effects vary, thus making pooling problematic.

Most of the social science articles reporting estimates from which impact estimates can be extracted provide many effect size estimates, whether due to multiple outcome measures on the same units (multiple end point studies), multiple treatment and multiple estimates using different model specifications or estimation methods, or variations on these themes. However, to maintain the independence assumption for meta-analysis, only one effect size per outcome-construct per study should be included in a single-level meta-analysis (Borenstein *et al*. 2009). In practice, there appear to be four alternative approaches to resolve this problem – to include all estimates, to drop some for which there can be a clear justification, to drop 'outliers' based on an arbitrary rule (for example, observations which are more than ± 2 standard deviations of the weighted mean) or to model the diversity (Borenstein *et al*. 2009).

Multivariate methods to model dependencies among effect estimates have also been proposed (Gleser and Olkin 1994, 2009, Kalaian and Raudenbusch 1996), but neither is the required information generally available unless the raw data can be accessed nor do these methods appear to have been implemented in commonly used software. Dropping outliers is arbitrary and can lose useful information. Lipsey and Wilson (2001) recommend averaging effects (p. 101), but whether this is appropriate when effect sizes remain heterogeneous remains unresolved. Where there are multiple statistical specifications of the effect size estimation in a single study, a more nuanced approach might be to choose a 'favoured' specification based on a risk of bias assessment, averaging (using inverse variance weights) across any remaining multiple outcome estimates. However, we also note that standard procedures for averaging effects neglect the likely correlation among effect size estimates within a study, which should be taken into account (Hedges and Olkin 1985, Rosenthal and Rubin 1986).

In the case of development studies, we are faced not only with diverse definitions and metrics of outcomes and treatment but also with diverse research designs, methods of analysis and reported parameters and statistics. Does it make sense to combine these diverse studies for meta-analysis? The next sections discuss these issues in more depth.

3.4. Effect size types

Summary outcome figures reported in published work come in many forms; for example, a simple comparison of a continuous outcome variable for a dichotomous treatment can be

reported as means of treatment and control groups, their number and standard deviations. The simplest standardisation for unpaired samples is indeed the difference between the means of the treated and the untreated divided by the standard deviation of the pooled data (Cohen's '*d*' – Cohen 1988), but this is held to be biased especially in small samples. Variations of Cohen's *d* derived in the context of small samples are Glass's 'Δ' and Hedges '*g*'. These types of outcome estimate are termed 'standardised mean difference' (SMD) outcomes.

Data reported in this form allow direct computation using the standard formulae. However, they may also be reported as the mean difference and its standard deviation or standard error of the difference, with or without the number of treatment and control units. Or they may be reported as the mean difference and a '*t*' value and so on. The most extensive free effect size calculator reports 30 different combinations of statistics that can be used to calculate an SMD effect size.[13]

Many social science studies in development report impacts in the form of regression statistics, particularly regression coefficients (betas – *b*). As is well known, *b* is the effect of one unit of the treatment variable on the dependent variable measured in its metric. Since these metrics can vary (height in metres is different from weight in kilogram and so forth) they need to be standardised. The treatment variable may also need to be standardised if it is not a dichotomy (0/1). Thus, some articles report standardised betas, which are the effect in standard deviations of a 1 standard deviation change in the treatment variable. This is problematic when using a dichotomous treatment indicator. Calculation of SMD effect sizes from regression studies requires information on outcome means and sample sizes by group, together with pooled standard deviations. Unfortunately, this information is often not reported.

Meta-analysis is generally conducted on summary outcome indicators such as those reported in the final publication, rather than through re-analysis of the raw data; notwithstanding the problems that this causes, this arises because access to the raw data is often impossible, especially for studies conducted in the past, and re-analysis (pure replication) can be extremely costly if the data are poorly documented and or have a complex structure. Nevertheless, this is acknowledged to be unfortunate; for example Chalmers *et al*. (2002) finish their brief history of research synthesis by urging that 'the future history of research synthesis should be based increasingly on the creation of publicly accessible archives of raw data' (p. 32). A compromise is for authors to provide more extensive reporting of statistics so that they can be used in meta-analysis. In particular, reporting of the standard deviation of the pooled data should be standard in development studies journals.[14] In an age of such low costs of electronic data storage and access, there can be no satisfactory reasons for not providing more extensive reports (King 2007).

Since studies, which are to be pooled, may not report effect sizes in equivalent forms, it may be desirable to translate the effect size into a common measure before undertaking any synthesis analysis. Several texts provide translation formulae (Lipsey and Wilson 2001, Ellis 2010).

4. Meta-analysis

Having arrived at a set of effect size estimates, it is possible to pool them in a meta-analysis. Combined analysis of quantitative results of research from different studies is an obvious way to synthesise research findings, and has a long history (Chalmers *et al*. 2002, Petticrew and Roberts 2006, pp. 192–193). The recent growth in meta-analysis dates from Glass (1976) and Smith and Glass (1977); the latter synthesised the effects of different

psychotherapies from 'controlled studies' on a standardised (by the standard deviation of the control group – Glass's Δ) mean difference in 'any outcome variable [that] the researcher chose to measure' (Smith and Glass 1977, p. 753). Glass and Smith (1979) synthesised studies on the effects of class size on achievement again including a number of different outcome variables and 'uncontrolled' as well as 'controlled' studies. The article compared the effects estimated from different designs (pp. 14–16).[15]

In addition to the effect size statistics (depending on the effect size metric used), their variance, number of units and an indicator of the study, variables representing characteristics of the studies may be extracted to be used to account for variations in effect sizes between studies, if sub-group analysis or meta-regression are to be undertaken. When the included studies give rise to different effect size statistics, which are to be transformed into a common metric, an indicator of the type of effect size statistic will also be needed, and pre-processing of the effect sizes to undertake this transformation using the standard equations and methods referred to above will be needed prior to the meta-analysis itself.

There have been many critics of meta-analysis (for example, Eysenck 1978, Shapiro 1994, Feinstein 1995, Berk and Freedman 2003).[16] The standard literature on meta-analysis is based on data from an experimental design using a common treatment with the same effect size metric based on means, response ratios and odds- or risk-ratios; estimates derived from multivariate methods, typically from studies using quasi-experimental data, can be included in meta-analysis, but, as noted above, because the different studies will often not have the same sets of covariates, the interpretation of results becomes more problematic (Becker and Wu 2007). Indeed, there are many authoritative critiques of meta-analysis based on quasi-experimental data (Egger *et al*. 1998), even though some of the early meta-analysis was quite cavalier about the inclusion criteria, for example Smith and Glass (1977). Some critiques are a priori – for example Shapiro (1994) proposed that given the low quality and high risk of bias of non-randomised studies, 'the meta-analysis of published non-experimental data should be abandoned' (p. 777). Others, however, have more nuanced positions, suggesting that the legitimacy of including effect sizes based on observational data in meta-analysis depends on their quality and the plausibility of controlling for biases. Controversies around the inclusion of studies using quasi-experimental data gave rise to the argument of the 'MOOSE' declaration on reporting (Stroup *et al*. 2000), who state that 'standards of reporting must be maintained to allow proper evaluation of the quality and completeness of meta-analyses' (p. 2012). Although doubts remain about the use of effect sizes from quasi-experimental data in meta-analysis (Jones 1992), Petticrew and Roberts (2006) seem to conclude that this is inevitable, but should be done with care (p. 207) (see also Smith and Egger 1999).[17]

4.1. Meta-analysis of experimental data

The simplest case is where all included studies have the same effect size statistic and are included only once. This situation does not guarantee freedom from bias because RCTs in practice do not live up to the idealised claims made on their behalf, even in the medical arena (see Petryna 2009 for an insightful discussion). Meta-analysis then computes an overall effect size, its confidence intervals and generally provides a forest plot to assist diagnosis. It will also conduct tests of the homogeneity of the estimates; if the individual studies give a wide range of estimated effects and or have large differences in their variability as judged by a Z-test (if there are only two sub-groups), Q-test or I statistics (Borenstein *et al*. 2009), an overall mean estimate derived from all the included studies will be considered unreliable, and attempts will be made to group the studies. Ideally, sub-groups should be

grouped a priori – say by intervention type if more than one intervention is present in the included studies. As elsewhere, statistical and substantive significance of the effects are not necessarily the same, and attention should be paid to effect sizes of different sub-groups and to whether it seems reasonable to identify causes of these differences through meta-regression.

Fixed effects meta-analysis calculates a pooled effect assuming there is a single 'true' common effect and that differences between studies are due to sampling errors. In contrast, random effects meta-analysis assumes that there are differences in 'true' effects due to differences in populations, implementation and so on, and estimates an 'average effect' and its variability (Riley *et al*. 2011). The choice between a fixed or random effects model is based on a priori reasoning, and in development studies, the random effects model is likely to be appropriate.

4.2. Meta-analysis of quasi-experimental estimates

The inclusion in a meta-analysis of effect sizes based on quasi-experimental designs creates additional methodological sources of heterogeneity. For example, the effect sizes estimated from PSM are based on the treatment effect for those individuals that receive the treatment (average treatment effect on the treated, ATET). Effect sizes computed from RDD and IV yield local treatment effects; in the case of RDD, these are local to those individuals at the margin of the cut-off point (local average treatment effect at the discontinuity, LATED); for IV, these are those individuals for which the instrument induces a change in the treatment status (local average treatment effect, LATE). Finally, effect sizes computed from the results of regression-based approaches, including those based on RCTs with perfect compliance, yield the impact of the programme on the whole sample population (average treatment effect, ATE).[18]

In some cases, it may be possible to convert treatment effects into a common measure (Bloom 2006).[19] Where this is not possible, two issues should be considered for each study before meta-analysis: firstly, whether we might expect heterogeneous effects of the interventions across the sample population, and secondly, to what extent the individuals for whom the treatment is estimated approximate a random draw of the sample population. Pooling effect sizes from different study designs is particularly problematic in instances when the participation process leads to participants being less representative of the sample population or when the allocation rule can influence the impact of the programme (for example, by selecting the most motivated individuals). This is analogous to the case of RCTs with imperfect compliance, where intention-to-treat (ITT) and other treatment effects estimate the impact of the programme (or of the assignment into the programme) for a different sub-group of the sample population (Green *et al*. 2011). When there are serious concerns about either treatment homogeneity or representativeness, pooling may only be appropriate across comparable treatment effects.[20]

4.3. Bias detection in meta-analysis

Since the social science studies in development included in meta-analysis are generally of questionable validity and readily vulnerable to systematic biases, it is important to be able to identify and if possible to control for these remaining biases. In the mainstream meta-analysis literature based on experimental data, it has long been recognised that there are biases in reporting results (Dickersin *et al*. 1987, Egger *et al*. 1997); common biases are to publish (find and include in meta-analysis) positive results, statistically significant

results and publications in English. Failure to report particular outcomes or entire studies with negative or no significant effects (also known as the 'file-drawer problem'; Rosenthal 1979) has more recently been seen as a research ethics problem leading to the establishment of registration and documentation of all trials[21] so that a fuller picture of all research on a topic is available.[22]

Given the studies included in a meta-analysis, some bias can be detected from the observation that the precision of effect sizes is inversely related to the number of units included in the study; larger studies are more likely to detect a significant effect leading to publication, but among smaller studies, only those with positive significant effects are likely to be put forward and accepted for publication. Suppose there is a small positive effect of the intervention, then large studies are likely to find a significant positive effect and be published. Among smaller studies with the same expected mean effect size, there will be a greater scatter around the true mean and also greater variability, so that even among those studies which find a positive impact and those that report a lower impact, there will be more scatter that are not statistically significant. This insight is displayed graphically in funnel plots (Light and Pillemer 1984, Egger *et al*. 1997). As is common practice, funnel plots assessing effect size against precision of effect (rather than sample size) are preferred, since statistical power reflects magnitude of effect as well as sample size (Green *et al*. 2011); this is especially relevant for quasi-experimental studies in development, which may be based on large sample observational data sets like national household surveys.

Publication bias of this type can also result from data mining or the process of developing hypotheses after the empirical results are known (which has gained the acronym of Hypothesising After the Results are Known, HARKing). Statistical tests have been developed based on the same insight (Egger, Harbord and Peters tests). Another approach has been to identify the 'fail-safe' N or number of additional studies with negative results that would be necessary to increase the P-value for the meta-analysis to above 0.05 (Cooper 1979, Rosenthal 1979, Orwin 1983, Becker 2006).

An approach to compensating for the file-drawer problem has been proposed by Duval and Tweedie (2000). The hypothesised missing studies are first trimmed and a new pooled estimate was derived from the now symmetric set of studies, before replacing both the trimmed studies and their missing counterparts and computing a final pooled effect size. Another approach is to model the selection of studies (Sutton 2009). Peters *et al*. (2007) found considerable variability in the effects of different trim and fill methods, but suggest that it can be used as a form of sensitivity analysis; Sutton (2009) concluded that publication (and related) biases remain 'a difficult problem to deal with' (p. 448). One may suggest that naive use of these methods can give an air of spurious reliability to a study.

5. Conclusion

We have seen that conducting meta-analysis of studies that rely on quasi-experimental designs, which have high levels of heterogeneity within and between studies and diverse conceptual framings, is a challenging task. This task is not made easier by the fact that many studies report insufficient details on their research design, method of analysis, descriptive statistics and impact estimates and their variability, complicating effect size calculations.

The discussion has highlighted two reasons for proceeding with caution when pooling effect sizes from quasi-experiments – risk of bias and methodological sources of heterogeneity. The former is only a problem if the studies are subject to high risk of bias. Methodological sources of heterogeneity in results arise due to allocation rules, which

impact the external validity of the effect estimate, and the effect of covariates in regression analysis.

In addition, the likely prevalence of researcher bias (also known as researcher allegiance), and any tendency to publish positive and (statistically) significant results suggest that if all actual studies on a given topic had been published, average effect sizes may well have been smaller, and possibly no more statistically significant. Together, it seems likely that the meta-analyses of effects in development studies are particularly vulnerable to systematic positive bias, deriving from the closeness of development studies to funder, activist, political or advocacy interests together with the usual publication biases. Under these circumstances, meta-analysis risks inflating statistical significance by combining poor-quality studies each of which at best yields only marginally significant results, and are vulnerable to unknown, but likely positive, biases which then inflate both effect sizes and confidence levels.

A sanguine view of this situation might argue the case for pooling and extensive subgroup analysis or meta-regression, provided sufficient studies with statistical homogeneity among at least some groups of estimations, together with analysis of publication bias. We can conclude from the discussion in this article that pooling the effect sizes calculated from studies of diverse methodological designs needs to be undertaken carefully because of the potential threats to internal and external validity and the dangers of compounding confounded estimates (Egger *et al*. 1998, Petticrew and Roberts 2006, pp. 204–205).

Notes

1. For example, the Campbell Collaboration has established an International Development Coordinating Group: http://www.campbellcollaboration.org/international_development/index.php.
2. For example, the leading medical text on systematic review and meta-analysis (Green *et al*. 2011) recommends, where possible, to include only experimental studies in meta-analysis. Only in the case of interventions to which experimentation is not applicable, do the authors acknowledge that non-randomised studies could be included in a meta-analysis. In contrast, Rubin (1974) and Angrist *et al*. (1996) provide good reasons to think that rigorously designed non-randomised quasi-experimental studies yield appropriate causal inference; see also Deaton (2010). Benson and Hartz (2000) and Concato *et al*. (2000) report similar effects from observational and RCT studies.
3. This framework should not be taken to suggest we endorse a universal 'hierarchy of methods' (Vandenbroucke 1989, White 2009); Duvendack *et al*. (2011) proposed this framework to screen a very large number of observational studies.
4. Interrupted time series (ITS) are often considered a high-quality method of analysis (see Shadish *et al*. 2002). They are not commonly used in development programme evaluation and are outside the scope of this article.
5. Internal validity assessment should include risk of biased point estimate and variance, the latter including 'unit of analysis errors' (Green *et al*. 2011) as well as heteroscedasticity in the case of regression-based studies.
6. For example, EPHPP (n.d.), EPOC (n.d.), NICE (2009), Valentine and Cooper (2008). For an early survey, see Deeks *et al*. (2003, Chapter 5); see also Petticrew and Roberts (2006).
7. See Salsburg (2001) for an entertaining introduction to the history of significance testing and its role in statistics. Ziliak and McCloskey (2008) provide a discussion in relation to economics.
8. There is some debate about the role of power calculations ex-post to claim that a study was 'under-powered', using the sample statistics. Since the sample statistics are just that, not a sample of samples, the use of estimated difference and its variability for the particular sample size begs the question of how representative the sample statistics in the particular case were (see Ellis 2010, pp. 58–59 on the 'perils of post-hoc power analyses').
9. Effect sizes computation might be desirable even though, due to heterogeneity or lack of independence within observations, meta-analysis is not possible. In combination with ex-post

power calculations this can reveal that a study is low powered, but this does not mean that the estimated effect size can be relied on (Ellis 2010, pp. 58–61); it could be wildly wrong.

10. However, it is recognised that randomisation may not achieve covariate balance, so tests of balance are necessary, and statistical control for unbalanced covariates is generally deemed necessary. Obviously, control for unbalanced covariates can only be attempted for observed potential confounders (Hansen and Bowers 2008).
11. For example, a *t*-test of differences in means of a continuous variable can be reported as *N*, mean and standard deviation of treatment and control groups, or as '*t*', '*F*'. '*p*' or '*z*' value together with total or sub-group sample sizes.
12. Duvendack *et al*. (2011) illustrate that studies of the impact of microfinance can be assessed in terms of business investments, or anthropometry, health or education. The former is presumably instrumental to the later (probably through net incomes), whereas the latter are both intrinsic to (constitutive of) human well-being, and instrumental.
13. See http://gemini.gmu.edu/cebcp/EffectSizeCalculator/d/means-and-standard-deviations.html.
14. Full reporting of variance–covariance matrices would also enable appropriate calculation of adjusted effect sizes from regression analysis (Becker and Wu 2007).
15. In neither the area of the effects of psychotherapy nor that of class size has a systematic review ended controversy (Hedges and Stock 1983, Wampold *et al.* 2000).
16. Whereas some of them are related with the meta-analysis of data from quasi-experimental studies (for example, Shapiro 1994), others such as Berk and Freedman (2003) highlight that the statistical validity assumptions underpinning the validity of meta-analysis are in most cases unfeasible.
17. It seems that some adepts of systematic reviews see its limitations: 'Are we investing too heavily in an excessively precise concept of some overall treatment effect instead of more closely examining the heterogeneity of findings, their nature, the biological explanation of such heterogeneity, and what it really means for decision making?' (Jenicek 2006, p. 3).
18. Nonetheless, regression-based approaches assume that conditional on covariates, the ATET is not different from the ATE on the total population.
19. Bloom (2006) provides guidance on how to transform among some treatment effects. However, these transformations also require information which may be seldomly reported.
20. Riley *et al*. (2011) argue for the use of 'prediction intervals' calculated by adjusting the random effects confidence interval by the between-study variance (the Tau-statistic). The prediction interval is therefore wider than the random effects confidence interval (itself wider than that for the fixed effect), and is interpreted as an estimate of the 'likely effect in an individual setting' (Riley *et al*. 2011, p. 964).
21. For example, the SWHO International Clinical Trials Registry Platform – http://www.who.int/ictrp/en/.
22. The International Initiative for Impact Evaluation (3ie) is proposing a registry for impact evaluations in development.

References

Angrist, J., Imbens, G., and Rubin, D., 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91 (434), 444–455.

Becker, B.J., 2006. Failsafe N or file-drawer number. *In*: H.R. Rothstein, A.J. Sutton, and M. Borenstein, eds. *Publication bias in meta-analysis: prevention, assessment and adjustments*. Chichester: Wiley.

Becker, B.J. and Wu, M.-J., 2007. The synthesis of regression slopes in meta-analysis. *Statistical science*, 22 (3), 414–429.

Benson, K. and Hartz, A.J., 2000. A comparison of observational studies and randomized, controlled trials. *New England journal of medicine*, 342 (25), 1878–1886.

Berk, R. and Freedman, D., 2003. Statistical assumptions as empirical commitments. *In*: T. Blomberg and S. Cohen, eds. *Law, punishment and social control: essays in honor of Sheldon Messinger*. New York: Aldine de Gruyter, 235–254.

Bloom, H., 2006. *The core analytics of randomised experiments for social research* [online], MDRC Working Papers on Research Methodology. Available from: http://www.mdrc.org/publications/437/full.pdf [Accessed 24 July 2012].

Borenstein, M., *et al*., 2009. *Introduction to meta-analysis*. Chichester: Wiley.
Coalition for Evidence-Based Policy, 2010. *Checklist for reviewing a randomized controlled trial of a social program or project, to assess whether it produced valid evidence* [online]. Available from: http://coalition4evidence.org/wordpress/wp-content/uploads/Checklist-For-Reviewing-a-RCT-Jan10.pdf [Accessed 24 July 2012].
Chalmers, I., Hedges, L.V., and Cooper, H., 2002. A brief history of research synthesis. *Evaluation and the health professions*, 25 (1), 12–37.
Cohen, J., 1988. *Statistical power analysis for the behavioural sciences*. 2nd ed. Hillsdale, NJ: Erlbaum Associates.
Colliver, J.A., Kucera, K., and Verhulst, S.J., 2008. Meta-analysis of quasi-experimental research: are systematic narrative reviews indicated? *Medical education*, 42 (9), 858–865.
Concato, J., Shah, N., and Horwitz, R.I., 2000. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342, 1887–1892.
Cook, T., Shadish, W., and Wong, V., 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *Journal of policy analysis and management*, 27 (4), 724–750.
Cooper, H., 1979. Statistically combining independent studies: a meta-analysis of sex differences in conformity research. *Journal of personality and social psychology*, 37 (1), 131–146.
Cooper, H.M. and Hedges, L.V., 1994. *The handbook of research synthesis*. New York: Russell Sage Foundation.
Cumming, G., 2012. *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Taylor & Francis.
Deaton, A., 2010. Instruments, randomization, and learning about development. *Journal of economic literature*, 48 (2), 424–455.
Deeks, J., *et al*., 2003. Evaluating non-randomised intervention studies. *Health technology assessment*, 7 (27), 1–173.
Dickersin, K., *et al*., 1987. Publication bias and clinical trials. *Controlled clinical trials*, 8 (4), 343–353.
Duflo, E., Glennerster, R., and Kremer, M., 2007. Using randomization in development economics research: a toolkit. *Handbook of development economics*, 4, 3895–3962.
Duval, S. and Tweedie, R., 2000. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56 (2), 455–463.
Duvendack, M., *et al*., 2011. *What is the evidence of the impact of microfinance on the well-being of poor people?* London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
Effective Practice and Organisation of Care Group (EPOC), n.d. *Suggested risk of bias criteria for EPOC reviews* [online]. Available from: http://epocoslo.cochrane.org/sites/epocoslo.cochrane.org/files/uploads/Suggested%20risk%20of%20bias%20criteria%20for%20EPOC%20reviews.pdf [Accessed 7 March 2012].
Effective Public Health Practice Project (EPHPP), n.d. *Quality assessment tool for quantitative studies* [online]. Available from: http://www.ephpp.ca/PDF/Quality%20Assessment%20Tool_2010_2.pdf [Accessed 7 March 2012]
Egger, M., Schneider, M., and Smith, G.D., 1998. Meta-analysis spurious precision? Meta-analysis of observational studies. *British medical journal*, 316, 140–144.
Egger, M., *et al*., 1997. Bias in meta-analysis detected by a simple, graphical test. *British medical journal*, 315 (7109), 629–634.
Ellis, P.D., 2010. *The essential guide to effect sizes: statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.
Eysenck, H.J., 1978. An exercise in mega-silliness. *American psychologist*, 33 (5), 517–517.
Eysenck, H.J., 1984. Meta-analysis: an abuse of research integration. *The journal of special education*, 18 (1), 41–59.
Feinstein, A.R., 1995. Meta-analysis: statistical alchemy for the 21st century. *Journal of clinical epidemiology*, 48 (1), 71–79.
Field, A., 2009. *Discovering statistics using SPSS*. 3rd ed. London: Sage.
Fisher, R.A., 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.
Glass, G.V., 1976. Primary, secondary and meta-analysis of research. *Educational researcher*, 10, 3–8.
Glass, G.V. and Smith, M.L., 1979. Meta-analysis of research on class size and achievement. *Educational evaluation and policy analysis*, 1 (1), 2–16.

Glazerman, S., Levy, D.M., and Myers, D., 2003. Nonexperimental versus experimental estimates of earnings impacts. *The Annals of the American Academy of Political and Social Science*, 589 (1), 63–93.

Gleser, I.J. and Olkin, I., 1994. Stochastically dependent effect sizes. *In*: H. Cooper and L.V. Hedges, eds. *The handbook of research synthesis*. New York: Russell Sage Foundation, 339–355.

Gleser, I.J. and Olkin, I., 2009. Stochastically dependent effect sizes. *In*: H. Cooper and L.V. Hedges, eds. *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation.

Green, S., *et al*., 2011. *In*: J.P.T. Higgins and S. Green, eds. *Cochrane handbook for systematic reviews of interventions version 5.1.0 (updated March 2011)* [online]. Available from: www.cochrane-handbook.org [Accessed 24 July 2012].

Grissom, R.J. and Kim, J.J., 2012. *Effect size for research*. London: Routledge.

Hansen, B.B. and Bowers, J., 2008. Covariate balance in simple, stratified and clustered comparative studies. *Statistical science*, 23 (2), 219–236.

Hansen, H., Klejntrup, N., and Andersen, O., 2011. *A comparison of model-based and design-based impact evaluations of interventions in developing countries* [online], FOI Working Paper 2011/16. Available from: http://okonomi.foi.dk/workingpapers/WPpdf/WP2011/WP_2011_16_model_vs_design.pdf [Accessed 24 July 2012].

Hedges, L.V. and Olkin, I., 1985. *Statistical methods for meta-analysis* [online], Academic Press. Available from: http://www.jameslindlibrary.org/illustrating/records/statistical-methods-for-meta-analysis/title_pages [Accessed 24 July 2012].

Hedges, L.V. and Stock, W., 1983. The effects of class size: an examination of rival hypotheses. *American educational research journal*, 20, 63–85.

Heinsman, D.T. and Shadish, W.R., 1996. Assignment methods in experimentation: when do non-randomized experiments approximate the answers from randomized experiments? *Psychological methods*, 1, 154–169.

Jenicek, M., 2006. *Méta-analyse en médecine: the first book on systematic reviews in medicine* [online]. Available from: http://www.jameslindlibrary.org/illustrating/articles/meta-analyse-en-medecine-the-first-book-on-systematic-reviews [Accessed 24 July 2012].

Jones, D.R., 1992. Meta-analysis of observational epidemiological studies: a review. *Journal of the royal society of medicine*, 85 (3), 165–168.

Kalaian, H.A. and Raudenbusch, S.W., 1996. A multivariate linear model for meta-analysis. *Psychological methods*, 1 (3), 227–235.

Keef, S.P. and Roberts, L.A., 2004. The meta-analysis of partial effect sizes. *The British journal of mathematical and statistical psychology*, 57 (1), 97–129.

King, G., 2007. An introduction to the dataverse network as an infrastructure for data sharing. *Sociological methods and research*, 36 (2), 173–199.

Light, R.J. and Pillemer, D.S., 1984. *Summing up: the science of reviewing research*. Cambridge, MA: Harvard University Press.

Lipsey, M.W. and Wilson, D.B., 1993. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American psychologist*, 48, 1181–1209.

Lipsey, M.W. and Wilson, D.B., 2001. *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

McGrath, R.E. and Meyer, G.J., 2006. When effect sizes disagree: the case of r and d. *Psychological methods*, 11 (4), 386–401.

Morgan, S. and Winship, C., 2007. *Counterfactual and causal inference: methods and principles for social research*. New York: Cambridge University Press.

National Institute for Health and Clinical Excellence (NICE), 2009. Quality appraisal checklist – quantitative intervention studies [online]. *In*: *Methods for the development of NICE public health guidance*. Available from: http://www.nice.org.uk/media/2FB/53/PHMethodsManual110509.pdf [Accessed 24 July 2012].

Nickerson, R.S., 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, 5 (2), 241–301.

Orwin, R.G., 1983. A fail-safe *N* for effect size in meta-analysis. *Journal of educational statistics*, 8 (2), 157–159.

Peters, J.L., *et al*., 2007. Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in medicine*, 26, 4544–4562.

Petryna, A., 2009. *When experiments travel: clinical trials and the global search for human subjects*. Princeton, NJ: Princeton University Press.

Petticrew, M. and Roberts, H., 2006. *Systematic reviews in the social sciences: a practical guide*. Oxford: Blackwell Publishing.
Raudenbusch, S.W., 2009. Analyzing effect sizes: random-effect models. *In*: H. Cooper, L.V. Hedges, and J.C. Valentine, eds. *The handbook of research synthesis and meta-analysis*. New York: Russell Sage, 295–316.
Ravallion, M., 2007. Evaluating anti-poverty programmes. *Handbook of development economics*, 4, 3787–3846.
Riley, R., Higgins, J., and Deeks, J., 2011. Interpretation of random effects meta-analysis. *British medical journal*, 342, 549.
Rosenthal, R., 1979. The "file drawer" problem and tolerance for null results. *Psychological bulletin*, 86, 38–641.
Rosenthal, R., 1991. *Meta-analytic procedures for social research*. 2nd ed. (original 1984). London: Sage.
Rosenthal, R. and Rubin, D.B., 1986. Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological bulletin*, 99, 400–406.
Rubin, D., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66 (5), 688–701.
Salsburg, D., 2001. *The lady tasting tea: how statistics revolutionised science in the twentieth century*. New York: W. H. Freeman/Holt.
Sen, A.K., 1999. *Development as freedom*. Oxford: Oxford University Press.
Shadish, W., Cook, T., and Campbell, D., 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Brooks/Cole Cengage Learning.
Shapiro, S., 1994. Meta-analysis/shmeta-analysis. *American journal of epidemiology*, 140 (9), 771–778.
Sharpe, D., 1997. Of apples and oranges, file drawers and garbage: why validity issues in meta-analysis will not go away. *Clinical psychology review*, 17 (8), 881–901.
Slavin, R.E., 1986. Best-evidence synthesis: an alternative to meta-analytic and traditional reviews. *Educational researcher*, 15 (9), 5–11.
Smith, G.D. and Egger, M., 1999. Meta-analyses of observational data should be done with due care. *British medical journal*, 318 (7175), 256.
Smith, M.L. and Glass, G.V., 1977. Meta-analysis of psychotherapy outcome studies. *American psychologist*, 32 (9), 752–760.
Stroup, D., *et al*., 2000. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *Journal of the American medical association*, 283 (15), 2008–2012.
Sutton, A.J., 2009. Publication bias. *In*: H.M. Cooper, L.V. Hedges, and J.C. Valentine, eds. *Handbook of replication studies in the social and behavioural sciences*. New York: Russell Sage, 435–452.
Sutton, A.J., *et al*., 1998. Systematic reviews of trials and other studies. *Health technology assessment*, 2 (19), 1–310.
Valentine, J. and Cooper, H., 2008. A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: the study design and implementation assessment device. *Psychological methods*, 13 (2), 130–149.
Vandenbroucke, J.P., 1989. Is there a hierarchy of methods in clinical research? *Journal of molecular medicine*, 67 (10), 515–517.
Wampold, B.E., Ahn, H.-N., and Kim, D.-M., 2000. Meta-analysis in the social sciences: a useful way to make sense of a series of findings from a large number of studies. *Asia Pacific education review*, 1 (1), 67–74.
White, H., 2009. *Some reflections on current debates in impact evaluation* [online]. New Delhi, 3ie Working Paper No 1. Available from: http://www.3ieimpact.org/media/filer/2012/05/07/Working_Paper_1.pdf [Accessed 24 July 2012].
Wilcox, R.R., 2008. Sample size and statistical power. *In*: A.M. Nezu and C.M. Nezu, eds. *Evidence based outcome research: a practical guide to conducting randomized controlled trials for psychosocial interventions*. New York: Oxford University Press, 123–134.
Wilcox, R.R., 2012. *Introduction to robust estimation and hypothesis testing*. 3rd ed. San Diego, CA: Academic Press.
Ziliak, S.T. and McCloskey, D.N., 2008. *The cult of statistical significance: how the standard error cost us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.